Modeling Space in Historical Texts

Gregory, Ian N.; Donaldson, Christopher; Hardie, Andrew; Rayson, Paul

Modelling space in historical texts

Ian Gregory, Chris Donaldson, Andrew Hardie and Paul Rayson

The texts used as sources in the humanities tend to be rich in information about place and space. Much of the information about place in such sources comes directly from the place-names mentioned in them, however, some comes indirectly through more ambiguous, descriptive geographical references, as in the phrases 'near the hills' or 'beyond the town.' In either case, the spatial information can be difficult for a human reader to conceptualise effectively. Even when a reader is familiar with the area described, identifying the location of the places to which that source refers and conceptualising how those places relate to one another can be difficult; when the reader is unfamiliar with the area it is all but impossible. Beyond this, gaining an in-depth understanding the way that different places are represented with a text often requires the places to be represented or modelled in non-textual form. One solution to this problem is to map the places to which the source refers. The effectiveness of this approach has been explored elsewhere (see, for example, Moretti, 1998). However, manually mapping textual information is time-consuming and the resulting maps are inflexible and, therefore, have only limited analytical potential.

Digital technologies present new opportunities for modelling, analysing and interpreting the geographical information contained within texts. This chapter will explore some of these opportunities, focussing particularly on the modelling, analysis and interpretation of geographical information in texts, and on the difficulties of integrating space in textual analysis. Although texts are a particularly rich source of geographical information, extracting this information from them is challenging. The approaches documented here, describe how place-names and the themes associated with them can be identified, visualised and analysed. Although these are applied to a specific corpus, they are inherently transferable, and can be readily adapted and applied to other sources.

The corpus utilised in this chapter is a digitised collection of historical texts about the English Lake District, a canonical literary and cultural landscape in the North West of England. The corpus contains eighty individual works, which range in date from 1622 to 1900 and together comprise over one and a half million words. It includes both famous and highly influential accounts of the Lake District, such as Thomas West's *Guide to the Lakes* (1778) and William Wordsworth's *Guide through the District of the Lakes* (5th ed., 1835), as well as a number of

lesser-known, but significant works, such as *Black's Shilling Guide to the English Lakes*, which appeared in no fewer than twenty-two editions between 1853 and 1900. It also includes a selection of famous literary texts, such as the 1622 edition of Michael Drayton's *Poly-Olbion*, and personal testimonials, such as the letters composed by the poet John Keats during his walking tour through the Lake District in 1818. Drawing on this corpus, in what follows we will demonstrate how space and the themes associated with it can be modelled and analysed – using a combination of geographic information systems (GIS) and other technologies. In the process, we will explain the procedures involved in translating textual data into a GIS format and evaluate the analytic opportunities that this affords.

Modelling features and their locations using GIS

Geographical information is conventionally represented digitally using a geographical information system (GIS). Effectively, GIS software takes conventional database functionality that allows data in tabular form to be stored, manipulated and queried, and combines this with a mapping system that, as well as mapping, also provides a range of other functionality for manipulating and querying information about location. To implement this a GIS requires two types of data: firstly, there are the tables of text or numbers that are typically thought of as 'data' and are familiar to users of database management systems such as Microsoft Access, MySql or even Microsoft Excel. In GIS parlance, these are referred to as *attribute data*. The mapping system requires that each row of data to be given a location representing the feature's location on the Earth's surface. These are referred to as *spatial data* and are based on co-ordinates. Simple features can be modelled using a single co-ordinate representing a point that models, for example, a village or a mountain top. Lines use a string of co-ordinates to represent, for example, roads or rivers. Areas, such as lakes or parishes, are represented using polygons, which are areas completely enclosed by one or more lines. Spatial data may also take the form of a *raster surface* in which the study area is broken down into a matrix of small regular pixels. Raster surfaces are often used to represent height. Taken together, the spatial data represent where features are located, while the attribute data provides information about what the features are (see, for example: Chrisman, 2002; Heywood, Cornelius and Carver, 2002; Gregory and Ell, 2007).

[Figure 1: Conventional GIS]

This data model is well suited to representing, visualising and analysing data from the Earth and environmental sciences and the social sciences. Figure 1 illustrates the use of a GIS *layer*

(as the combination of spatial and attribute data is called) to display historical census data representing population in and around the English Lake District in 1851. The attribute data are contained within the table, a fragment of which is shown in 1b. In this case the attribute data include information about the name of the parish and which district and county it is in, as well as statistical information on the total population in 1841 and 1851. The spatial data are the polygons that represent the historical parishes of the region.

This two-part data model has obvious merits as a tool for visualisation, but additionally it allows space to be used to query, analyse and integrate information (Gregory, Kemp and Mostern, 2003). Querying the data spatially allows the researchers to ask 'what is at this location?' and 'what is near this location?'; it moreover allows the results of more conventional queries of attribute data, such as 'what parishes have values greater than 4,000?', to be represented visually. Indeed, the shading on Figure 1a is in fact a response to a series of queries on the attribute data that have first selected polygons with a population of greater than 4,000 and shaded them with the darkest shading, then selected polygons with values between 2,001 and 4,000 and shaded them in a lighter shade, and so on. From an analytic perspective, additional approaches can be developed that make use of the proximity and distances between features, and thus enable us to ask questions about whether and why phenomena or events seem to happen in certain places and not in others. Integration through location allows additional data to be added to the model of the study area. As spatial data are represented using real-world co-ordinates that are either in latitude and longitude, or in a projection system such as British National Grid or Universal Transverse Mercator any dataset with spatial data can, in theory, be merged with any other dataset based on their geographical location.

Representing texts using GIS

GIS have shown themselves to be well suited to modelling tabular data for which precise locations can be determined. Examples include census and vital registration data (Gregory, 2008; Beveridge, 2014), land-use data (Cunfer, 2005) and economic data (Knowles and Healey, 2006). Archaeologists have also found GIS helpful for conducting landscape surveys and for representing data about excavations (Wheatley and Gillings, 2002; Conolly and Lake, 2006). Unfortunately most humanities scholars do not work with these types of structured, quantitative data. Instead, in some cases they will have relatively structured lists of place-names that they want to explore in map form, for example a list of places of publication taken

from a library catalogue or corpus metadata. More often, however, they will have corpora of unstructured texts, usually stored as one or more text files, which may include mark-up in formats such as XML (eXtensible Mark-up Language, see Hardie, 2014) which possibly follows a standard such as TEI (Text Encoding Initiative, see Barnard and Ide, 1997). While these texts may be known to contain place-names, which names and where they are in the text will usually be unknown. If we are to better understand the geographies within these types of sources the challenge is thus how to convert a text into a data model suitable for inclusion in GIS.

[Figure 2: Example of using a gazetteer]

The major challenge in converting a text into a format suitable for use in GIS is usually providing a co-ordinate that can be used as spatial data for each place-name. When a researcher already has a list of place-names, such as places of publication, converting this to GIS format is relatively straightforward. The easiest way to do this is to use a place-name gazetteer such as Geonames.[1] A gazetteer is effectively a database table that includes place-names and their co-ordinates (Southall, Mostern and Berman, 2011). A relational join can be used to attempt to automatically match all of the place-names from the researcher's list to the gazetteer. This will create a new table that adds the co-ordinates and any other relevant information from the gazetteer to the researcher's original data. The researcher needs to check the results, as relational joins often fail to match due to minor spelling variations. There may also be additional problems such as place-names that do not occur in the gazetteer, or ambiguous place-names that can refer to more than one place in the gazetteer. These will need some manual intervention from the researcher. Figure 2 shows an example of this process: an input table that contains a hypothetical list of place-names associated with a text is joined with a gazetteer to add co-ordinates. In one case, Ulverstone, the match has failed because of different spellings between the two tables; this would need to be rectified manually. It is also important to note that the gazetteer is likely to have many more place-names than the input table, but those which do not match any of the input data are not copied over to the output table.

Once we have a table where co-ordinates have been added to a list of place-names, importing this into GIS software should be a simple task. Any GIS software package should be able to

[1] See: http://www.geonames.org. Other gazetteers are available including: Getty Thesaurus of Geographic Names (http://www.getty.edu/research/tools/vocabularies/tgn) and the Ordnance Survey 1:50,000 Scale Gazetteer (http://www.ordnancesurvey.co.uk/business-and-government/products/50k-gazetteer.html).

take such as table of data and use the co-ordinates to *geo-reference*, as providing real-world co-ordinates is called, it. The result is a point layer which can be displayed as a dot map.

Where a user has an unstructured text containing place-names, an additional level of complexity is added. Converting this into GIS format now becomes a two-stage process, where first the place-names have to be identified and then they have to have co-ordinates allocated to them. Given a relatively small corpus, this can be done by reading through the whole text and manually identifying the place-names (Gregory and Cooper, 2009; Cooper and Gregory, 2011). They can either be copied and pasted into a table which can then be geo-referenced as described above, or they can be tagged using XML (as described below) to allow them to be extracted at a later stage. Manually identifies place-names gives high levels of accuracy, but is too labour intensive to be practical for all but small corpora.

Identifying place-names and allocating them to co-ordinates can be done automatically on larger corpora using a process called *geo-parsing* (Grover, et al., 2010). The first stage in this involves using techniques from natural language processing (NLP, see Jurafsky and Martin, 2008; Manning and Schütze, 1999 for introductions) to attempt to identify all of the place-names within the text or corpus in question. This takes advantage of the fact that place-names are proper nouns, or named entities, which can be identified and tagged using advanced NLP techniques. Geo-parsing techniques can also exploit contextual information to suggest whether the proper nouns are place-names as opposed to personal names or the names of organisations. This provides a list of 'candidate' list of words that are suspected of being place-names. The list of candidates is then matched to a gazetteer in the way described above. In the case of geo-parsed candidate place-names, the very fact of whether or not the candidate can be matched in the gazetteer is one factor that can help us decide whether that candidate is really a place-name. When a text is geo-parsed, information from the gazetteer is usually encoded back into the text using XML tags. The output from this process is thus the original text with added XML elements that identify place-names, with their co-ordinates and potentially other information as attributes.

[Figure 3: Geo-parser output]

Figure 3 shows an example of the output produced by the automated geo-parsing of West's (1778) *Guide to the Lakes* using the Edinburgh Geo-parser (Grover, et al., 2010). This

particular geo-parsing software identifies all candidate place-names using an 'enamex' tag.[2] Where these refer to place-names a range of additional information is stored as attributes, including the word number of the place-name within the text ('sw') and its longitude and latitude ('long' and 'lat'). Other information derived from the gazetteer includes: the type of place that it is, its gazetteer ID, and a standardised version of its spelling ('type', 'gazref', and 'name' respectively). A confidence score, calculated by the geo-parser to help disambiguate places with the same name ('conf'), is also included. Thus, in Figure 3, we can see that the two identified place-names which have been assigned latitudes and longitudes – Skiddaw and Helvellyn – are mountains (type="mtn"), have standardised spellings which are the same as their spelling within the text, and have been assigned latitudes and longitudes.

Although tools such as the Edinburgh Geo-parser enable us to geo-parse large corpora automatically, manual intervention will ensure a higher degree of accuracy. Place-name identification is a complex and subjective process, and there are multiple sources of geo-parsing error. Common examples include: failing to identify a proper noun and thus missing a place-name; wrongly identifying a personal name or other word as a place-name; giving the wrong co-ordinate to a place–name; spelling variations between the source text the gazetteer (including those caused by digitisation errors); and so on. As an example, take the following sentence: 'Travelling over the Raise, the Bishop of Carlisle paused to admire the view of Windermere before continuing on to Langdale'.[3] Carlisle is a city near the Lake District but, as 'Bishop of Carlisle' is a title, in most cases, but perhaps not all, we would not want to include this as a place-name. Windermere is both a lake and town. Whereas a human reader is likely to infer that the reference here must be to the lake, a computer is unlikely to be so subtle. In the centre of the Lake District there is a valley called Great Langdale, which runs parallel to a smaller valley called Little Langdale. Colloquially, the name 'Langdale' is used to refer to Great Langdale; however, there is an entirely separate but comparatively unknown valley called Langdale in North Yorkshire, east of the Lake District. A computer comparing candidate place-names from the text with names from a gazetteer is most likely to match Langdale to the last of these, whereas a human is likely to assume that it refers to Great Langdale. Finally, is 'the Raise' a place-name at all? It is likely to refer to 'Dunmail Raise', a pass that connects of Grasmere and Thirlmere, but it is highly unlikely that a geo-parser would identify it as this. Moreover, one might argue that because the word 'raise' is a

[2] In XML tags are enclosed by '<' and '>' symbols and the information to which each tag refers concludes with an end-tag ('</…>')
[3] This is an artificial sentence designed to illustrate the issues, not an actual quote.

generic, regional term for a pass, 'the Raise' should not be identified as a place-name at all. A counter-argument would be that, because it is spelt with a capital 'R', 'Raise' is being used as a proper noun and that should thus should be considered a place-name. In the context of a non-modern corpus, this argument is further complicated by the fact that, as we go back in time, English *common* nouns are more likely to receive the capitalisation that more modern English applies only to proper nouns. All this goes to show that the decision of what is and is not a place-name can be highly subjective and a researcher's definition may evolve as the research proceeds.

Errors such as these meant that, in their own assessment of the accuracy of the Edinburgh Geo-parser, Tobin et al (2010) found that 75% of place-names were correctly located when using a thirteen-million word corpus of official reports from the nineteenth and early twentieth centuries. Although this represents a good start, clearly this contains a lot of error; moreover the extent to which this will bias results and cause problems with an analysis is unclear. Additionally, when a different corpus is geo-parsed, it is not clear whether the results will be better or worse than this. One way to address this is through the use of *concordance geo-parsing* (Rupp et al., 2014). This is based on the idea that geo-parsing the entire corpus in one go is both unnecessary and difficult to correct. Instead, only the text that occurs near to a search-term of interest is geo-parsed. As an example, if we are interested in the term 'sublime' (which is a recurrent term in historical writing about the Lake District), we would first identify and extracted all of the instances of this word from the text along with their *concordance lines* (the text that occurs either side of them). In this case, we extract concordance lines containing 50 words to the left and 50 words to the right of the search-term to provide co-text for the geo-parser to work with. Extracting this is relatively simple using corpus analysis software such as AntConc or CQPweb (see Anthony, 2013 and Hardie, 2012 respectively). These concordance lines are geo-parsed using the Edinburgh Geo-parser. The results can then be examined for errors both by mapping and reading the concordance lines. Because there are far fewer words to explore than in the entire corpus, it is far easier to check these results. Any corrections that need to be made are then written to an updates file so that, when another search-term is used, these updates will be automatically applied as part of the geo-parsing process. In this way, the researcher can check and correct material manually, ideally starting with a search-term that is relatively rare in the corpus – in order to ensure that the amount of checking remains manageable – and working up to more common terms as the updates file develops. While this semi-automated process takes time, it greatly improves

accuracy as the researcher is now in control of the process. Additionally, it encourages the researcher to become familiar with the ways that place-names are used in the corpus and, therefore, to become aware of issues such as ‘Raise’ and ‘Bishop of Carlisle’ and to make their own decision about whether these are place-names that should be assigned co-ordinates and, if so, to which location each particular instance of this term should refer. For large corpora there is an additional advantage, since when billions of words are involved, the processing time of geo-parsing the entire corpus may be too long to be practical; thus it makes more sense only to geo-parse as required.

The output from the geo-parsing process is a text with additional XML mark-up that identifies place-names and gives their co-ordinates. The last stage is to convert this text into a GIS layer. This involves using a program that extracts information about every occurrence of a place-name from the XML and writes it to a table. At a minimum this table needs to have the place-name and its associated co-ordinates. To make it more usable, however, a range of other fields are also likely to be needed. One obvious source of information that is likely to be required is the other data about the place-name which, in figure 3, would be additional information from the attributes of the ‘enamex’ tag. Some additional textual information is also likely to be required. A word number to give the location of the place-name within the text might also be helpful. If concordance geo-parsing is used the search-term should also be used as well as the text of the concordance line itself, so that both place-name and search-term can be viewed in context. This presents some problems as in most databases, including those that underlie popular GIS software, text fields can only be a maximum of 255 characters long. Additional metadata that shows the source of the information may also be included.

[Figure 4: Example record]

Figure 4 shows one example of the columns of the table based on a concordance geo-parsing around the word ‘sublime’ in our Lake District corpus. The first four records, down to ‘title,’ are derived from the corpus metadata and give information about the text in which the place-name was found. The next five, down to the standardised place-name, are taken from the geo-parser output (other fields such as ‘type’ and ‘conf’ from Figure 3 could also have been included). The last six items are from the text itself, giving the place-name as it appears in the

text, the search-term, and the co-text found to the left and right of both of these. In this example co-text fields have been restricted to ten word tokens.[4]

[Figure 5: Sublime point layer]

Geo-parsing allows the geographies associated with themes within a particular text to be explored. Themes are identified using one or more key-words and their geography is established based on which place-names collocate with these key-words. Obviously concordance geo-parsing is well suited to this. Figure 5 shows an example of modelling the geography of word 'sublime' in the Lake District corpus using a point map of the type that can quickly be created once a text has been geo-referenced. The points on the map represent place-names that occur within 10 words of the word 'sublime.' This map is clearly only a first stage in understanding the geography of this theme: while we have managed to locate the place-names that collocate with our search-term, the map itself tells us little about the geographical collocation pattern, what is creating it, or how it changes over time or between different genres. Understanding the pattern thus requires further modelling and analysis.

Analytic modelling of space

[Figure 6: Density smoothed sublime]

Figure 6 moves towards a more analytic modelling of a point pattern to enable us to understand the geographical distribution of place-name instances. It is based on the data from Figure 5, showing the geography of 'sublime'; however, rather than using points, the pattern has been smoothed to indicate which places have the highest number of points nearby. This is a process known as *density smoothing* (Lloyd, 2007: ch. 7). Technically, in GIS terminology, what is happening here is that we are converting from a point layer to a raster surface made up of small pixels. The values for the pixels are calculated from the distance from each pixel to its surrounding points, with nearer points given a higher weighting. The density smoothed map shows us a much clearer pattern than the points in Figure 5. It is easily seen that there are some clear clusters of places that tended to be described as sublime. Working roughly clockwise, these include Keswick, Borrowdale (south of Keswick), Ullswater, Windermere, Coniston, and the western fells from Sca Fell to the Pillar. Some caution must be used in interpreting this map. Geo-parsing represents all locations using points including the large

[4] Corpus linguists refer to individual instances of words in a text as '[word] tokens'. In most corpus software, a punctuation symbol is counted as a separate token; thus "red, white and blue" is five tokens (four words and a comma).

lakes such as Ullswater, Windermere and Coniston. The Ullswater point lies in the crook of the lake, approximately half-way along its length. Windermere and Coniston are both the names of villages as well as lakes, and the points representing them lie in the locations of the villages, on the east side of Windermere and the north-west of Coniston. All of these clusters may, therefore, refer to larger areas than they appear to be on the map so the maps must be interpreted accordingly.

Figures 5 and 6 illustrate the basics of modelling how a theme varies over space. They are, however, only crude abstractions which provide basic descriptions of the patterns. Geographical Text Analysis (GTA) (Murrieta-Flores et al., 2014) is a set of techniques that allows us to go further. It allows us to identify locations where place-name instances cluster together, and find what words collocate with the place-names in this cluster. Clusters can then be compared to see how different places vary. The spatial pattern of a particular search-term can also be compared with the overall distribution of place-name instances in the corpus to see where the search-term occurs more or less than would be expected given this background population. This enables the researcher to evaluate the extent to which a pattern such as Figure 6 is caused by people visiting and writing about places such as Keswick more than other places. Geographical patterns can also be compared, for example to see whether two search-terms have similar geographies; to investigate whether two authors or genres follow similar geographies; or to compare the geographies of place-name instances that occur in writing from different time periods. Other sources of data can also be integrated into the analysis to explore the relationship between place-name instances and, for example, heights, population density, roads or railway stations. Gregory and Donaldson (in press) and Donaldson, Gregory and Murrieta-Flores (2015) provide some examples of this for work on the Lake District.

GTA approaches are types of ‘macro-reading’ in which we move from the bare text to abstract summaries such as maps, graphs or statistics. These are used to identify patterns, and we then re-engage with the text in an attempt to explain the patterns found. (The preservation of concordance text in the GIS table is of great practical use in helping us get back to reading the underlying language of the text or corpus.) An alternative approach is to use Place-Centred Reading (Hastings, Gregory and Atkinson, 2015). This is a more traditional reading approach, but rather than reading in a linear manner, the researcher’s reading is based around place-names. The first stage is to identify one or more place-names of interest and search the corpus for these. The text around the place-names is then studied closely and other relevant

place-names are identified. The researcher then builds up a qualitative understanding of everything that has been written about a particular area.

Conclusions

This essay has explored the some of the basic challenges of how space can be modelled from digital texts. The major challenge in doing this lies with how to identify place-names within a large corpus and allocate them to a co-ordinate. Geo-parsing presents a solution to this issue, and our refinement of concordance geo-parsing offers further advantages in terms of accuracy and checking. Geo-parsing will not always be necessary. With smaller texts, perhaps up to a few tens of thousands of words, it may be possible, or even desirable, to simply identify place-names manually by reading through the whole text. Alternatively, a researcher may not need to identify all place-names within the text but may instead only be interested in locations that are associated with a text as a whole, for example, place of publication. Whatever the nature of the source, the text is abstracted to a table which contains place-names, their co-ordinates, and other information about the place derived from the text or the gazetteer. The co-ordinates are then used to create a GIS point layer. More sophisticated analysis can then be used to describe, visualise and understand the patterns further.

One major limitation with these approaches is that computers require representations of geography that is very precise and unambiguous. Humanities sources that consist of human language-in-use – that is, a discourse which, for us as humans, is inseparable from the social and cultural context in which it is situated – are rarely precise and unambiguous, and this causes a range of issues. For example, gazetteers tend to provide a point location for place-names. While this is suitable for features such as towns or mountains, it is far less suitable for others such as lakes, valleys or rivers. Lakes or rivers could be georeferenced using polygons or lines respectively, but it is in fact not entirely clear that these represent an improvement. If a text talks about “walking along the banks of Ullswater” or “we crossed the River Greta” they are not referring to the entire lake or river; therefore is representing the entire feature really an improvement on using a single point? If we do use lines and polygons, a second issue is how we can compare these with the points that represent other features. There is also the issue that we mentioned above, of place-names such as ‘Windermere’ which can refer to different features depending on context.

A second limitation is that rather than modelling space, the approaches described above actually model place-names and are less effective with less precise representations. One

example of this is if a writer says, for example, "the road from Keswick to Ambleside is scenic", an automated technique will wrongly identify that it is the two towns that are scenic, rather than the road between them. More intractable problems arise when place-names are not used or are only implied. When a writer uses phrases such as "a magnificent view of the hills" or "we arrived into town" it may, or may not, be obvious to a reader what time or place they are talking about; however, a computer is not able to work out the reference of such phrases reliably, and it is unlikely that technical developments will enable them to do so in the foreseeable future. Geography may also be taken as a more generic concept, for example, how writers represent upland areas, valleys or forest may be of interest, rather than the specifics of named places. Some of the ideas outlined above might help in understanding this, for example, by identifying which words collocate with terms such as 'mountain' or 'forest' but these would require different forms of visualisation than mapping.

Identifying and extracting the geographical information in a text or corpus is only the first stage in modelling its spatial characteristics, but is frequently the hardest and most time consuming. Once this information has been extracted, further modelling allows it to be analysed in more depth. This can involve using techniques from GIS, spatial analysis, time-series analysis, corpus linguistics (Adolphs, 2006; McEnery and Hardie, 2012) or combinations of these. These approaches provide macro-reading summaries of the text and the spatial and potentially temporal patterns associated with particular themes. It can also involve using spatial references to allow the texts to be read in a non-linear way, such that a reader can read everything the corpus says about a particular place or area, perhaps at a specific time or in relation to a particular theme. Crucially, however, all stages of the research process, the researcher needs to be wary and critical of what the map or other visualisation is showing. The types of maps produced by GIS provide an excellent way of crudely summarising spatial patterns, however, for the researcher to truly understand the patterns that emerge, and the reasons why they exist, requires significant further work in exploring the parts of the text from which the points on the maps were derived. In some cases this will lead to the map or graph being refined. Thus, a map in a GIS is very different to a map created by manual cartography. In manual cartography the map is in many ways the end product. In GIS a 'map', in the form of a layer of geo-referenced data, is produced early in the research process, and it is explored, corrected, queried and enhanced throughout the research process. The final output is not simply a static map or maps in a publication; the actual intellectual product we are working towards is the analysis or argument that the map helps to illustrate –

an analysis or argument that was, in turn, derived from the process of creating and refining the map.

Acknowledgements: The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant "Spatial Humanities: Texts, GIS, places" (agreement number 283850). We are also grateful to Dr. Claire Grover (University of Edinburgh) for her assistance in providing us with the Edinburgh Geo-parser and to Dr. David Cooper (Manchester Metropolitan University) for his work on creating the Lake District corpus.

Adolphs, S. 2006. *Introducing Electronic Text Analysis*. London: Routledge.

Anthony, L., 2013. Developing AntConc for a new generation of corpus linguists. In: *Proceedings of the Corpus Linguistics Conference (CL 2013)*. Lancaster, UK, 22-26 July 2013.

Baines, E., 1829. *A Companion to the Lakes of Cumberland, Westmoreland, and Lancashire: In a descriptive account of a family tour and excursions on horseback and on foot.* London: Simpkin and Marshall.

Barber, S. 1892. *Beneath Helvellyn's Shade. Notes and Sketches in the Valley of Wythburn*. London: E. Stock.

Barnard, D.T. and Ide, N.M., 1997. The Text Encoding Initiative: Flexible and extensible document encoding. *Journal of the Association for Information Science and Technology*, 48, pp. 622-628.

Beveridge, A.A. The development, persistence, and change of racial segregation in U.S. urban areas, 1880-2010. In: I.N. Gregory and A. Geddes, eds. 2014. *Towards Spatial Humanities: Historical GIS and Spatial History*. Bloomington, IN: Indiana University Press.

Conolly, J. and Lake, M., 2006, *Geographical Information Systems in Archaeology*. Cambridge: Cambridge University Press.

Cooke, C., 1827. *The Tourist's and Traveller's Companion to the Lakes*. London: Sherwood, Jones and Co.

Cooper, D. and Gregory, I.N., 2011. Mapping the English Lake District: A literary GIS. *Transactions of the Institute of British Geographers*, 36, pp. 89-108.

Cunfer, G. 2005. *On the Great Plains: Agriculture and environment*. College Station: Texas A & M University Press.

Chrisman, N., 2002. *Exploring Geographic Information Systems.* 2nd ed. New York: John Wiley & Sons.

Dalton, J., 1755. *A Descriptive Poem, Addressed to Two Ladies, at their Return from Viewing the Mines near Whitehaven.* London: J. J. Rivington.

Donaldson, C., Gregory, I. and Murrieta-Flores, P. (2015) Mapping 'Wordsworthshire': A GIS study of literary tourism in Victorian England. *Journal of Victorian Culture*, 20, pp. 287-307

Gilpin, W., 1786. *Observations, Relative Chiefly to Picturesque Beauty.* London: R. Blamire

Gregory, I.N., 2008. Different places, different stories: Infant mortality decline in England & Wales, 1851-1911. *Annals of the Association of American Geographers*, 98, pp. 773-794.

Gregory, I.N., Kemp, K. and Mostern R., 2003. Geographical Information and historical research: Current progress and future directions. *History and Computing*, 13, pp. 7-21.

Gregory, I.N. and Ell, P.S., 2007. *Historical GIS: Techniques, methodologies and scholarship*. Cambridge: Cambridge University Press.

Gregory, I.N. and Cooper D., 2009. Thomas Gray, Samuel Taylor Coleridge and Geographical Information Systems: A Literary GIS of Two Lake District Tours. *International Journal of Humanities and Arts Computing*, 3, pp. 61-84.

Gregory, I. and Donaldson C., in press. "Using geographical technologies to understand Lake District literature." In D. Cooper, C. Donaldson and P. Murrieta-Flores P, eds. *Literary Mapping in the Digital Age*. Ashgate

Gregory, I., Cooper, D., Hardie, A., and Rayson, P., in press, 2015. Spatializing and analysing digital texts: Corpora, GIS and places. In D.J. Bodenhamer, J. Corrigan and T.M. Harris, eds. *Spatial Narratives and Deep Maps.* : Bloomington: IN, Indiana University Press. pp. 150-178.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J., 2010. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A*, 368, pp. 3875-3889.

Hardie, A., 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17, pp. 380–409.

Hardie, A., 2014. Modest XML for corpora: Not a standard but a suggestion. *ICAME Journal*, 38, pp. 73-103.

Hastings, S., Gregory, I.N. and Atkinson, P., 2015. Explaining geographical variations in English rural infant mortality decline using place-centred reading. *Historical Methods*, 48, pp. 128-140.

Heywood, I., Cornelius, S. and Carver S., 2002. *An Introduction to Geographical Information Systems.* 2nd ed. Harlow, Essex: Prentice Hall.

Jurafsky, D. and Martin, J.H. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. Harlow, Essex: Prentice Hall.

Knowles, A.K. and Healey, R.G., 2006. Geography, timing, and technology: A GIS-based analysis of Pennsylvania's iron industry, 1825-1875. *Journal of Economic History*, 66, pp. 608-634.

Lloyd, C.D., 2007. *Local Models for Spatial Analysis*. Boca Raton: FL, CRC Press.

Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press

McEnery, A.M. and Hardie, A., 2012. *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Moretti, F., 1998. *Atlas of the European Novel 1800-1900.* London: Verso.

Murrieta-Flores, P., Baron, A., Gregory, I., Hardie, A. and Rayson P., 2014. Automatically analysing large texts in a GIS environment: The Registrar General's reports and cholera in the nineteenth century. *Transactions in GIS*

Radcliffe, A., *A Journey Made in the Summer of 1794*. London: G. G. and J. Robinson

Rupp, C.J., Rayson, P., Gregory, I., Hardie, A., Joulain, A., and Hartmann, D., in press, 2015. Dealing with heterogeneous big data when geoparsing historical corpora. In: *IEEE Conference on Big Data*, Bethesda: MD, 27 Oct 2014

Southall, H., Mostern R. and Berman M.L., 2011. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5, pp. 127-145.

Tobin, R., Grover, C., Byrne, K., Reid, J. and Walsh J., 2010. Evaluation of georeferencing. *Proceedings of the 6th Workshop on Geographic Information Retrieval*. Zurich: ACM.

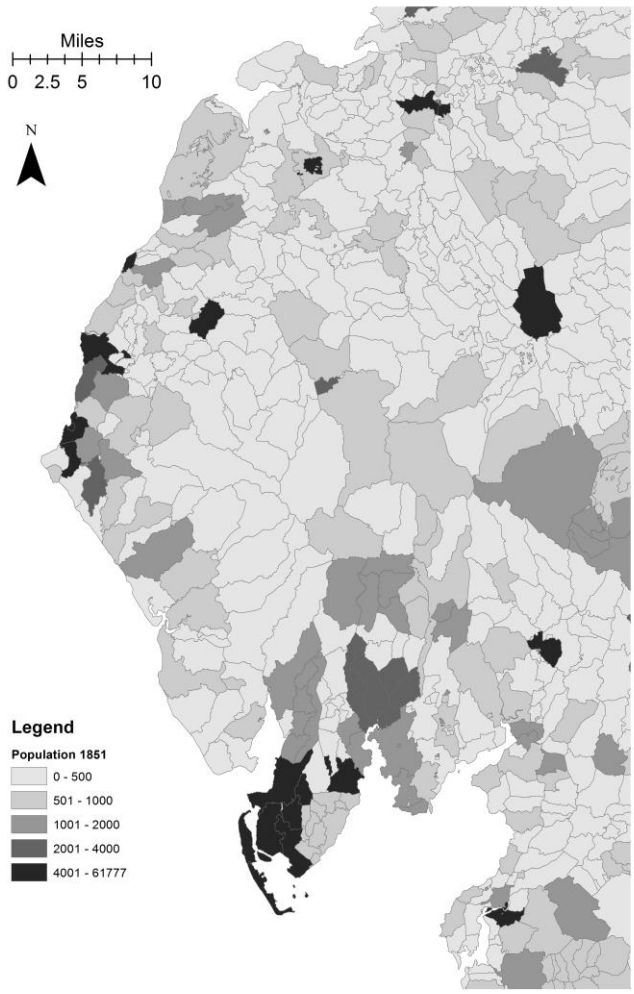
West, T., 1778. *A Guide to the Lakes: Dedicated to the lovers of landscape studies.* London: Richardson and Urquhart.

Wheatley, D. and Gillings, M., 2002. *Spatial Technology and Archaeology: The archaeological applications of GIS*. London: Taylor & Francis.

Wilkinson, T. 1824. *Tours to the British Mountains, with the Descriptive Poems of Lowther*. London: Taylor & Hessey.

Wordsworth, W. 1822. *A Description of the Scenery of the Lakes in the North of England*. London: Longman & Co.

Young A., 1770. *Six Months' Tour Through the North of England.* London: W. Strahan.



a) Spatial data

County	District	Parish	POP_1841	POP_1851
Cumberland	Penrith	Addingham	259	244
Cumberland	Wigton	Aikton	190	219
Cumberland	Penrith	Ainstable	501	524
Lancashire	Ulverstone	Aldingham	907	968
Westmorland	West Ward	Askham	193	179

b) Fragment of attribute data

Figure 1: A traditional GIS data model of 1851 census data. In (a) the spatial data are the polygons whose boundaries are shown and whose areas are shaded using values taken from the attribute data, a fragment of which is show in (b).

Input Table		
Text ID	**Date**	**Place Name**
1	1769	London
2	1771	Kendal
3	1773	Keswick
4	1784	Ulverstone
5	1790	London

Gazetteer		
Place Name	**Longitude**	**Latitude**
London	51.5	-0.1
Kendal	54.3	-2.8
Keswick	54.6	-3.1
Ulverston	54.2	-3.1
Carlisle	54.9	-2.9
Whitehaven	54.6	-3.6

Output table				
Text ID	**Date**	**Place Name**	**Longitude**	**Latitude**
1	1769	London	51.5	-0.1
2	1771	Kendal	54.3	-2.8
3	1773	Keswick	54.6	-3.1
4	1784	Ulverstone		
5	1790	London	51.5	-0.1

Figure 2: A simplified example of adding co-ordinates to an input table using a gazetteer.

To render the tour more agreeable, the company should be provided with a telescope, for viewing the fronts and summits of the inaccessible rocks, and the distant country, from the tops of the high mountains <*enamex sw="w14842" long="-3.123" lat="54.655" type="mtn" gazref="unlock:11284755" name="Skiddaw" conf="2.4"*>Skiddaw<*/enamex*> and <*enamex sw="w14854" long="-3.012" lat="54.530" type="mtn" gazref="unlock:11169753" name="Helvellyn" conf="2.4"*>Helvellyn<*/enamex*>.

Figure 3: An example of output from the Edinburgh Geo-Parser. This is a fragment of text from Thomas West's Guide. The 'enamex' tags encode the geographical information as described in the text.

Field	Explanation	Example Value
FileId	Unique ID for the text the place-name came from	34
Author	…of this text	Anon.-T. Ostell (pub.)
Year	…the text was published	1804
Title	…of the text	Observations, Chiefly Lithological, Made in a Five Weeks' Tour
WordNo	Location within the text where the place-name occurs	w25383
Latitude		54.545899
Longitude		-3.275492
GazRef	Identifies the gazetteer record used to geo-parse this place-name	unlock:11094751
StName	Standardised spelling of the place-name	Buttermere
LPlCotext	The co-text to the left of the place-name	, the sign of the Salmon . The scenery about
PlName	The place-name as it occurs in the text	Buttermere
RPlCotext	The co-text to the right of the place-name	is truly sublime and august . On a promontory to
LSTCotext	The co-text to the left of the search-term	of the Salmon . The scenery about Buttermere is truly
SearchTerm		sublime
RSTCotext	The co-text to the right of the search-term	and august . On a promontory to the east of

Figure 4: Example fields used to convert a geo-parsed text file to a table suitable for GIS and a sample record. This is based on concordance geo-parsing around the search-term ‘sublime’ in our Lake District corpus.

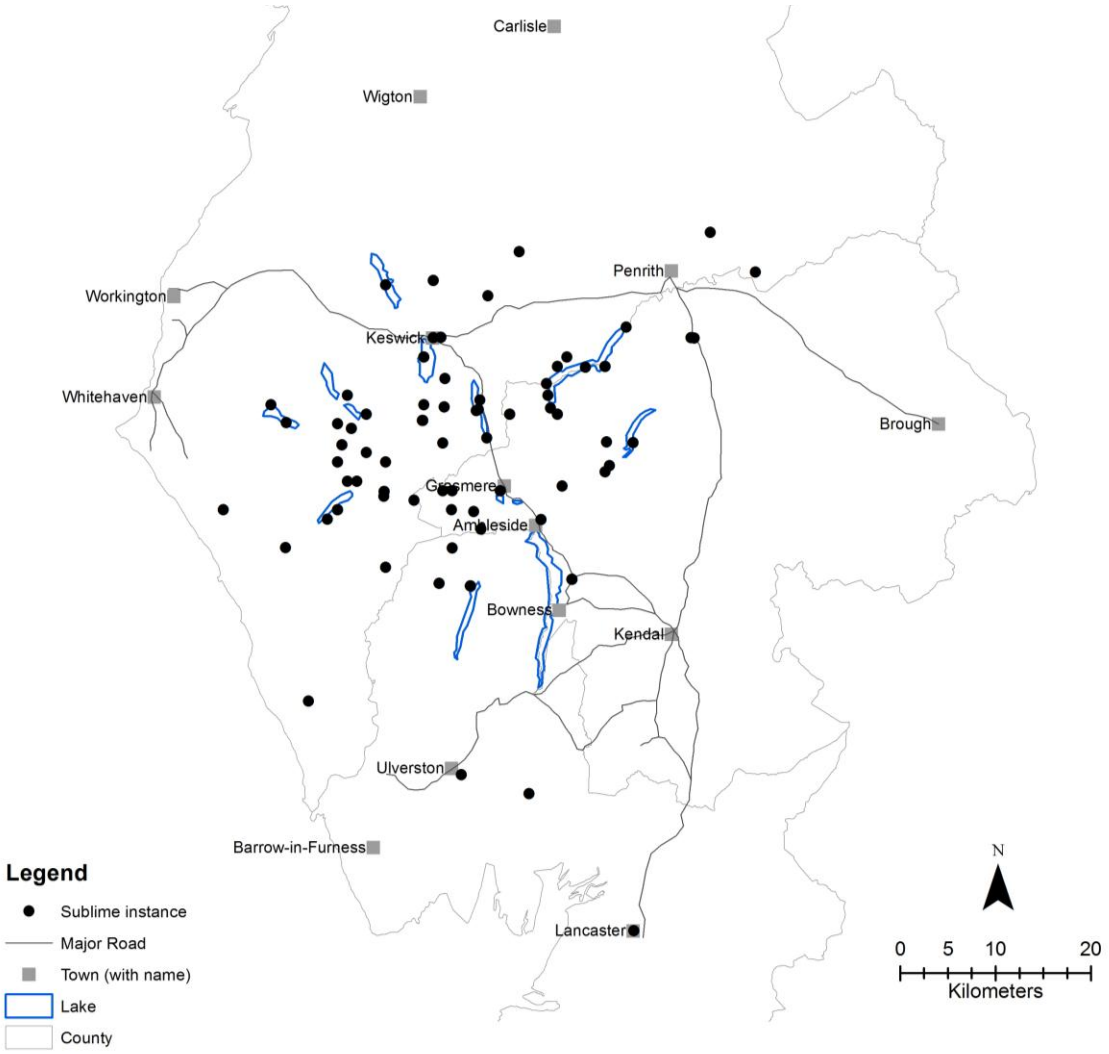


Figure 5: Instances of 'sublime' in the Lake District corpus.

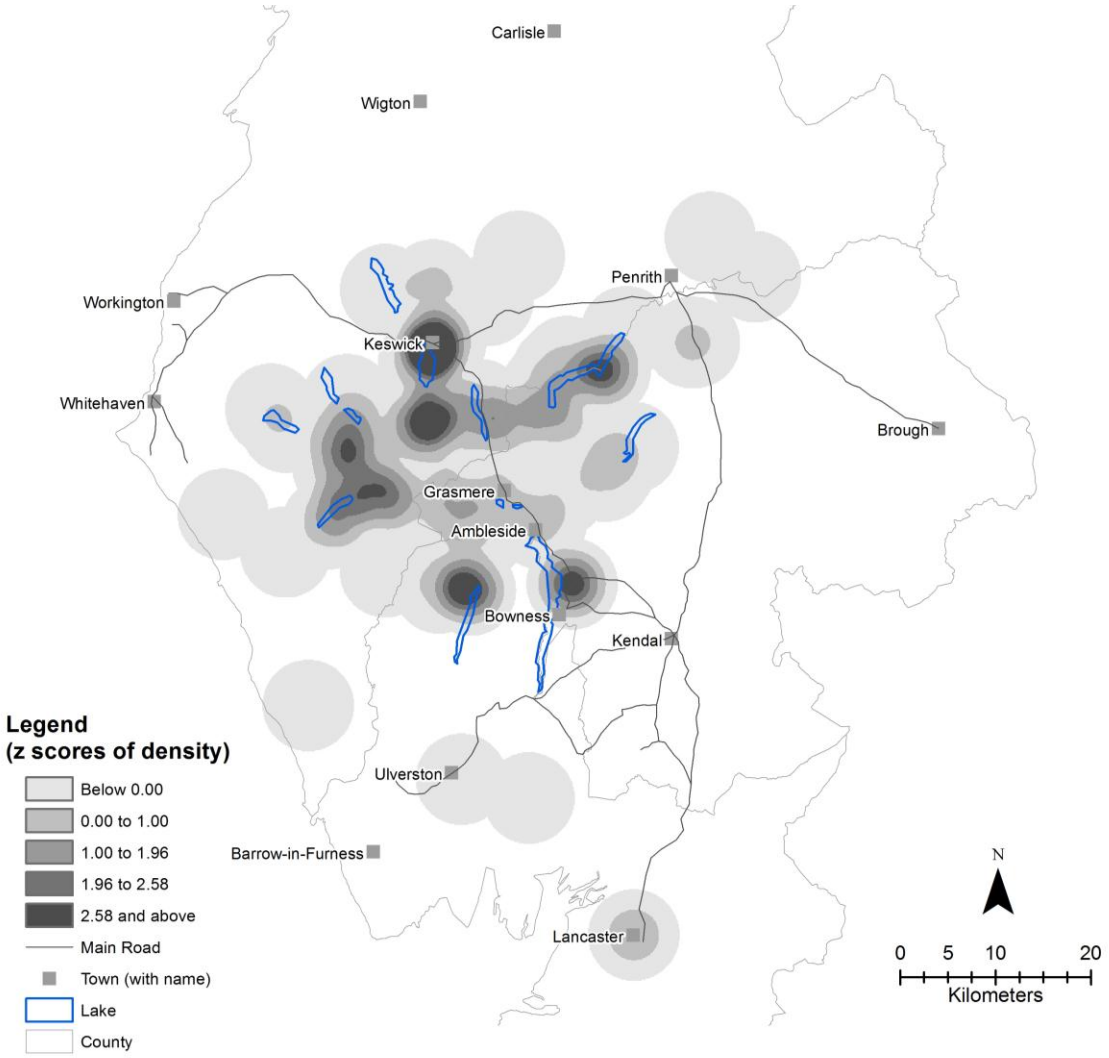


Figure 6: Density smoothed map of 'sublime' instances. Class intervals are based on critical values of z-scores of positive density values so 0.00 is the mean, 1.00 is one standard deviation over the mean, 1.96 is the 5% threshold (two-tailed) and 2.58 is the 1% threshold.